



INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386



Hahn, Waschulzik:

On the Use of Local RBF Networks to Approximate Multivalued Functions and Relations

Sonderforschungsbereich 386, Paper 116 (1998)

Online unter: <http://epub.ub.uni-muenchen.de/>

Projektpartner



On the Use of Local RBF Networks to Approximate Multivalued Functions and Relations

Klaus Hahn

Institute of Biomathematics and Biometrics of the GSF - National Research Center for Environment and Health, Postfach 1129, D-85758 Oberschleißheim, Germany

E-mail: hahn@gsf.de

Thomas Waschulzik

Technology Center Informatics, Intelligent Systems, FB 3,
University of Bremen, Bibliothekstr. 1, D-28359 Bremen, Germany

Abstract

A connectionist model made up of a combination of RBF networks is proposed; the model decomposes multivalued dependencies into local single valued functions; theory and applications are presented.

1 Introduction

We present a new network structure which is modelled according to the "implicit function theorem" [1]. It roughly states that multivalued functions and relations, which can be described by the zeros of an implicit function, can locally be represented by single valued functions. In this network, the local functions are realized by feedforward networks (RBF). They are incorporated into a global network by a symmetric topological encoding of the in- and output spaces and by a product of error functions. The latter represent separated classes of local functions. Via a least square training of the global network it is decided which one of the local networks generalizes best in a special region. This optimal network is then used for the local generalization of the multivalued function.

This construction performs interpolation and classification tasks on the same automatic control level as standard RBF networks. Moreover, it uses the generalization quality and the transparent parametrization of feedforward networks in the treatment of multivalued functions and relations. A regularization (smoothing) term is included in the model. Such networks are relevant in image smoothing, where discontinuities like edges pose a severe problem to numerous filters [2]. In addition, the use of least square training allows a precise treatment of the regression problem. Moreover in the field of non unique inverse problems, like in spectral analysis - where parameters of the unknown spectrum are to be determined -, or like in medical image analysis - where e.g. from the noisy image of a tumour its sharp contour (the regression curve) shall be reconstructed -, one frequently has to deal with relations.

2 The global network and its local RBF constituents

In this section, the essentials of the construction will be developed. The (local) single valued mappings used, are for notational simplicity one dimensional:

$\mathcal{R} \rightarrow \mathcal{R}$. In contrast to such single valued functions, which can be described e.g. as mappings $x \mapsto y$, multivalued functions and relations map x not only to a single value y , but also to several discrete y values or even to intervals. Thus for relations, we cannot expect to find an expression like $y = \text{network}(x, p)$ which gives all y -values depending on x , after some training process determining the parameters p of the generalizing network. To find the y -values given x for a relation, we therefore use the least square error function which was the basis of the training process and calculate from this function, $\text{error}(x, y)$, the zeros along the y axes for a fixed x (the case "given y " is treated in a symmetrical way).

Though relations globally differ from functions, they can locally be described by functions $x \mapsto y$ or $y \mapsto x$, at least if they are not too pathological. Since no coordinate axis is discriminated the input and output spaces will be encoded in a symmetric way; i.e. we introduce a x-axis-RBF set $\{\Phi_i(x) | i = 1 \text{ to necessary number of radial basis functions to cover that region of the x-axis which is occupied by the orthogonal projection of the relation on } x\}$, and a y-axis-RBF set $\{O_j(y) | \dots\}$ in an analogous way. This encoding of the in- and output spaces can be called topological, since adjacent points are described (encoded) by RBF's with adjacent centres. A local function $x \mapsto y$ may then be described by the term:

$$\text{error}(x, y)_{x \mapsto y, j} = [O_j(y) - \text{sig}(\sum_i w_{ij} \Phi_i(x))]^2 \quad (1)$$

where the sigmoid, $\text{sig}(x) = 1/(1 + \exp[-4x])$, has scaling properties. The center of $O_j(y)$ localizes the neighborhood in the y space, where $x \mapsto y$ can be trained to the data via $\text{error}(x, y)_{x \mapsto y, j}$; where $O_j(y) \approx 0$ this neighborhood has reached its limits. Looking for the zeros of $\text{error}(x, y)_{x \mapsto y, j}$ after training we find two problems:

- a) there are two generalizing solutions, symmetric to the center of $O_j(y)$,
- b) to reduce the neighborhood not too much, we must use a broad function $O_j(y)$, with the consequence, that the valley of zeros is unpleasantly flat.

These problems can be circumvented by summation over (1)

$$\text{error}(x, y)_{x \mapsto y} = \sum_j [O_j(y) - \text{sig}(\sum_i w_{ij} \Phi_i(x))]^2 \quad (2)$$

This represents all local functions in the direction $x \mapsto y$, which are regarded in some multivalued interpolation problem. Equation (2) can be interpreted as an implicit function in the sense mentioned in the Introduction. To illustrate this Ansatz, we show in Fig. 1 an interpolation, where via (2) a multivalued function $x \mapsto y$ with crossing and discontinuities is approximated by the training points indicated.

The mappings $y \mapsto x$ are summarized by the symmetric expression

$$\text{error}(x, y)_{y \mapsto x} = \sum_k [\Phi_k(x) - \text{sig}(\sum_l w_{kl} O_l(y))]^2 \quad (3)$$

The advantages of (2) and (3), however, have a price: if the training to the data cannot be performed perfectly as is usually the case, the error surfaces (2)

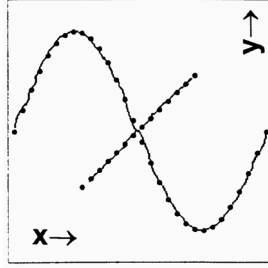


Fig. 1 A multivalued function $x \mapsto y$ is generalized by (2), (44 training points).

or (3) show no more perfect zeros as solution curves, but a course of minima somewhat above 0. The precise detection of this course is not trivial and will be discussed in section 3.

Up to now only multivalued functions can be treated. In order to generalize a relation (e.g. a spiral), a product of (2) and (3) is used:

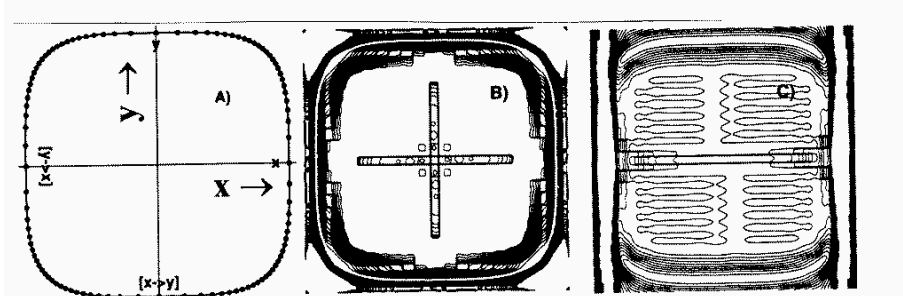


Fig. 2 The error surfaces of a relation described by four local functions

An example of a generalizing error function (4) - after training - is given in Fig. 2, panel B); the valley of minima follows the pattern "suggested" by the training points given in panel A). The error functions of the active local functions in the two factors of (4), in (2) and (3), are shown in panel C). As (2) generalizes the maps $x \mapsto y$ only, clear minima are present only at the top and bottom of panel C) (in panel A) the coordinate axes are given). To find the minima of (3) in C) just rotate panel C) by 90° . Panel A) additionally shows the continuous course of minima of panel B).

Summarizing: the error functions ["implicit functions"] (2), (3) and (4) define global networks by the course of minima ["zeros"]. These courses generalize multivalued functions ((2), (3)) and relations (4). The global networks are composed of RBF's ["local functions"] which generalize the data locally. As will be exemplified in Fig. 4, for some interpolation tasks, not only the networks are finally relevant, but also their extensions to the error surfaces.

3 Training process and ridge detection

In this section, the calculational procedure will be explained. The training points are the input, the course of minima in the error surface finally is the

output, which must be analyzed by some algorithm. The number of basis functions is given initially by the recipe: Choose the centres and widths of the Gaussians $\Phi_i(x)$ and $O_j(y)$ heuristically (in case of exact data: width of RBF's \approx center of RBF's distance \approx mean training point distance, in case of noisy data: width \approx center distance \approx standard deviation of training point density). For a relation the weights w_{ij} are obtained by a minimization of (4) plus a regularization term:

$$\min_w \left[\sum_{p=1}^{n_{training}} error(x_p, y_p) + \alpha \sum_{ij} (const + w_{ij})^2 \right] \quad (5)$$

where (x_p, y_p) are training points and α is the regularization strength. For Fig. 1 - Fig. 4 $\alpha \approx 0.05$ was used, for noisy data, $\alpha \approx 0.5$; $const = 0.5$ to achieve a flat error surface. Via the fit (5), the error function "decides" where (2) or (3) is generalizing best (builds a course close to zero), the w-matrix in the other term causes a bounded factor only. Therefore it is sufficient to use the same w-matrix in the factors (2) and (3) of (4). In contrast to a Spline interpolation, the training points need not to be ordered, the network generalizes to the nearest neighbours, sequential learning is possible. In this work a standard procedure of the NAG numerical library was used to perform the minimization. With respect to smoothness of the local RBF-networks as penalty term the weight-decay regularization was used [3]. Finally a fine tuning of the parameters can be performed via cross validation.

According to this training process, the error surface is minimized in the training points, the generalization extends these minima to smooth courses, which constitute the generalizing continua. A strict mathematical definition of courses or ridges is still an object of controverse debates [4]. However, there exists a numerical algorithm, which solves the problem of ridge calculation on a finite grid. To apply a variant of this watershed algorithm, $-error(x, y)$ is computed and discretized. In order to identify the pairs (x, y) which belong to the modelled multivalued function, the crest lines are extracted [5]. Although the original watershed algorithm detects only watershed and not crest lines in general, it can be modified by inserting artificial borders into the original image, before the watershed is computed. With this modified algorithm, it is possible to detect crest lines or ridges without adding artefacts. Fig. 3 and Fig. 4 demonstrate further applications of the described method.

4 Noisy data

In case of a quadratic error function, one can show for a continuous set of noisy training points that training of the weight matrix w for a standard RBF network (e.g. for the $x \mapsto y$ map) leads to a separation of the error function into a bias and a variance term. The minimum of the bias defines the network, which approximates the regression (e.g. $\langle y|x \rangle$) [3]. Similar reasoning leads for (2) and (3) to the same separation, the bias term of (2) is:

$$error_{x \rightarrow y}(x, \langle y|x \rangle) = \sum_j \left[\langle O_j|x \rangle - sig\left(\sum_i w_{ij} \Phi_i(x)\right) \right]^2 \quad (6)$$

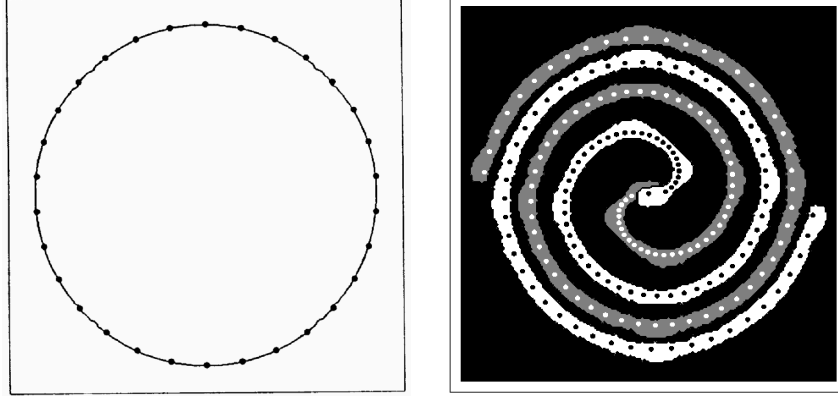


Fig. 3 A circle is defined by 30 points, the generalizing course is shown.

Fig. 4 The two spiral problem is treated by (4). Not only the courses are shown, but a three level colouring of the error surfaces.

where $p(y|x)$ denotes the conditional distribution of training points y for fixed x and

$$\langle f|x \rangle = \int f(y)p(y|x)dy \quad (7)$$

is a conditional mean of some function f .

In case of a Gaussian distribution $p(y|x) = e^{-(y-\langle y|x \rangle)^2 / (2\sigma(x)^2)} / \sqrt{2\pi\sigma(x)}$, we find for $O_j(y) = e^{-(y-m_j)^2 / (2s_j^2)}$:

$$\langle O_j|x \rangle = e^{-(\langle y|x \rangle - m_j)^2 / (2(s_j^2 + \sigma(x)^2))} s_j / \sqrt{s_j^2 + \sigma(x)^2} \quad (8)$$

For (3) the bias term is completely symmetric to (6). Since the course of (4) is defined by the minima of (2) or (3), in case of noisy data, (4) is replaced by the product of (6) and its symmetric counterpart. To exploit this modification, we must assume an approximate analytical form for the conditional point densities $p(y|x)$ and $p(x|y)$ (both densities include the unknown regression parameters $\langle y|x \rangle$, $\langle x|y \rangle$ and the standard deviations $\sigma(x), \sigma(y)$ as known). In regions of the plane, where (6) is active, the network approximates the regression $\langle y|x \rangle$, where the counterpart is active the regression $\langle x|y \rangle$. The training process is still performed by (5), which is free of distribution assumptions. In Fig. 5 and Fig. 6 applications of the described procedure are shown.

5 Summary and outlook

A new variant of a RBF network is proposed. It is applied to the approximation of multivalued functions and relations and seems to be a reasonable tool for generalization tasks, for non unique inverse problems and for special classification problems. The model generalizes exact and noisy data. In case of noisy data in the calculational procedure (not in training), assumptions about

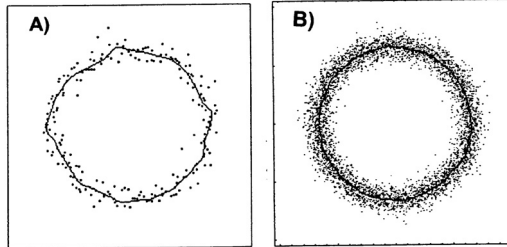


Fig. 5 The training points in panel A) are derived by a radial Gaussian distribution for 200 equidistant angles. The calculated curve is an estimator for the generating circular regression. In panel B) 5000 angles are used to demonstrate consistency.

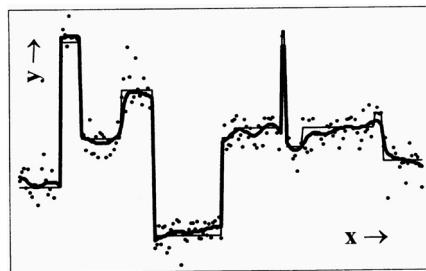


Fig. 6 The ability of (6) to approximate noisy discontinuous functions is demonstrated; thick line: (6), thin line: true regression, 200 training points are used [2].

the underlying point distributions enter. The network is defined for relations up to now by a symmetric Ansatz which originates in the use of two coordinate systems. In case of noisy data, however, the introduction of more coordinate systems might reduce the bias of this model. The model is presented in the plane only. Extensions to more dimensions are straightforward at least for the case of discontinuous image smoothing, Fig. 6. In case of true relations the construction of the model indicates how extensions to higher dimensions should be formulated.

References

- [1] R. Courant and F. Ford, Introduction to Calculus and Analysis, Springer-Verlag, Berlin Heidelberg New York, 1989.
- [2] C. K. Chu, I. Glad, F. Godtlielsen, J. S. Marron, Edge preserving smoothers for image processing, preprint, 1996.
- [3] C. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.
- [4] J. J. Koenderink and A. J. van Doorn, Image Structure, Invited Lecture, Mustererkennung 1997, 19. DAGM-Symposium, Braunschweig, Springer
- [5] L. Najan, R. Vaillant, Topological and geometric corners by watershed, in Hlavac, Sara (Eds.): CAIP 95 Proceedings, LNCS 970, Springer, 1995.